CLAIMS

1. A method for determining a measure of the level of expertise applicable to a given information data set, comprising the steps of:

(i) selecting, in respect of each of a plurality of predetermined levels of expertise, a representative sample set of information data sets;

(ii) determining, for each of said selected information data sets, the value of a metric indicative of the incidence, in a reference corpus of information, of terms comprised in the selected information data set; and

(iii) using the values of said metric determined in step (ii) to train an information classifier to identify, from a value of said metric calculated for the given information data set, at least one of said plurality of predetermined levels of expertise applicable to the given information data set.

2. A method as in Claim 1, wherein said metric comprises a combined measure of the incidence within an information data set of terms comprised in the information data set and of the incidence of each said term in the reference corpus.

3. A method as in Claim 1 or Claim 2, wherein at step (iii), training the classifier comprises:

(a) making distributions of normalised values of said metric for data sets in each of the representative sample sets selected at step (i); and

(b) for each of said predetermined levels of expertise, identifying from said distributions a corresponding range of normalised values of said metric.

4. A method as in any one of claims 1 to 3, wherein at step (iii), the trained classifier is arranged to determine a measure of the probability that a particular one of said predetermined levels of expertise is applicable to the information data set.

5. A method as in any one of the preceding claims, wherein determining a value for said metric comprises applying a stemming algorithm to stem terms comprised in a respective information data set and determining the incidence of the stemmed terms in the reference corpus.

6. A method as in any one of the preceding claims, wherein the reference corpus is provided with an interface for outputting the relative frequency of occurrence in the corpus of a term.

7. A method of accessing information data sets, stored in an information system, relevant to search criteria specifying an indication of a category of information to be accessed and to a specified indication of a predetermined level of expertise in respect of said category of information, the method comprising the steps of:

(i) selecting a training set of information data sets comprising, for each of a plurality of predetermined levels of expertise, a representative sample set of information data sets;

(ii) determining, for each data set in the training set, the value of a metric indicative of the incidence, in a reference corpus of information, of terms comprised in the training data set;

(iii) using the values of said metric determined in step (ii) to train an information classifier to identify at least one of said plurality of predetermined levels of expertise applicable to a given information data set;

(iv) applying an information searching algorithm to identify information data sets stored in said information system relevant to said specified category of information; and

(v) using the classifier trained at step (iii) to determine respective levels of expertise for information data sets identified at step (iv) and comparing the determined levels of expertise with the specified level of expertise to thereby select relevant information data sets.

8. An apparatus for determining a level of expertise applicable to an information data set, the level of expertise being selected from a plurality of predetermined levels of expertise, the apparatus comprising:

an input for receiving an information data set;

calculating means arranged with access to a reference corpus of information to calculate, for an information data set, the value of a metric indicative of the incidence, in the reference corpus, of terms comprised in the information data set;

a trainable classifier; and

training means for training said classifier to identify, using a training set of information data sets comprising, for each of said plurality of predetermined levels of

expertise, a representative sample set of information data sets and respective values of said metric, an applicable level of expertise selected from said plurality of predetermined levels of expertise for a received information data set;

wherein, in operation, on receipt of an information data set at said input, said calculating means are arranged to calculate a respective value for said metric and to input the calculated value to said trainable classifier, trained by said training means, to determine and output an indication of at least one of said plurality of predetermined levels of expertise applicable to said received information data set.

9. An apparatus as in Claim 8, wherein said metric comprises a combined measure of the incidence within an information data set of terms comprised in the information data set and of the incidence of each said term in the reference corpus.

10. An apparatus as in Claim 8 or Claim 9, wherein said training means are arranged to train said trainable classifier using the steps of:

(a) making distributions of normalised values of said metric for data sets in each of the representative sample sets; and

(b) for each of said predetermined levels of expertise, identifying from said distributions a corresponding range of normalised values of said metric.

11. An apparatus as in any one of claims 8 to 10, wherein said trainable classifier is arranged, after training by said training means, to determine a measure of the probability that a particular one of said plurality of predetermined levels of expertise is applicable to a received information data set.

12. An apparatus as in any one of claims 8 to 11, wherein said calculating means are arranged to calculate a value for said metric by applying a stemming algorithm to stem terms of a respective information data set and by determining the relative incidence of the stemmed terms in the reference corpus.

13. An information retrieval apparatus for accessing information data sets, stored in an information system, relevant to received search criteria specifying an indication of a category of information to be accessed and to a specified indication of a predetermined level of expertise in respect of said category of information, the apparatus comprising:

calculating means arranged with access to a reference corpus of information to calculate, for an information data set, the value of a metric indicative of the incidence, in the reference corpus, of terms comprised in the information data set;

a trainable classifier;

training means for training said classifier to identify, using a training set of information data sets comprising, for each of a plurality of predetermined levels of expertise, a representative sample set of information data sets and respective values of said metric, an applicable level of expertise selected from said plurality of predetermined levels of expertise for a given information data set;

searching means for identifying information data sets in said information system relevant to said specified category of information to be accessed; and

selecting means arranged to trigger said calculating means to calculate values of said metric for information data sets identified by said searching means, to input the values so calculated to said trainable classifier, trained by said training means, to determine and output respective applicable levels of expertise selected from said plurality of predetermined levels of expertise, and to select, for access, information data sets from those identified by said searching means having respectively determined levels of expertise that match said specified level of expertise.

14. An information retrieval apparatus for accessing information data sets, stored in an information system, relevant to received search criteria specifying an indication of a category of information to be accessed and to a specified indication of a predetermined level of expertise in respect of said category of information, the apparatus comprising:

calculating means arranged with access to a reference corpus of information to calculate, for an information data set, the value of a metric indicative of the incidence, in the reference corpus, of terms comprised in the information data set;

an information classifier, trained, using, for each of a plurality of predetermined levels of expertise, a representative sample set of training information data sets and respective values of said metric, to determine a level of expertise, selected from said plurality of predetermined levels of expertise, applicable to an information data set;

searching means for identifying information data sets in said information system relevant to said specified category of information to be accessed; and

selecting means arranged to trigger said calculating means to calculate values of said metric for information data sets identified by said searching means, to input the values so calculated to said information classifier to determine and output respective

applicable levels of expertise selected from said plurality of predetermined levels of expertise, and to select, for access, information data sets from those identified by said searching means having respectively determined levels of expertise that match said specified level of expertise.

5